Token-level Mean Switch Score (Top Tokens per Direction)